DOCUMENT RESUME

ED 102 205 TM 004 264

AUTHOR Hollenbeck, Albert R.; Slaby, Ronald G.
TITLE Measures of Reliability in Behavioral Observation: The Advantage of "Real Time" Data Acquisition.
SPONS AGENCY Washington Univ., Seattle. Child Development and Mental Retardation Center.
PUB DATE [74]
NOTE 21p.
AVAILABLE FROM Albert R. Hollenbeck, Department of Psychology, N.I-25, University of Washington, Seattle, Washington 98195

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Codification; *Digital Computers; Infant Behavior; Measurement Techniques; *Observation; Test Reliability; Video Tape Recordings
IDENTIFIERS Behavioral Observation Scoring System

ABSTRACT

Two observers who were using an electronic digital data acquisition system were spot checked for reliability at random times over a four month period. Between-and within-observer reliability was assessed for frequency, duration, and duration-per-event measures of four infant behaviors. The results confirmed the problem of observer drift--the fluctuations of scores across sessions--for the frequency and duration-per-event measures. In contrast, the "real time" duration scores were stable across sessions, indicating the robustness of this measure of behavior. (Author)

ED102205

MEASURES OF RELIABILITY IN BEHAVIORAL OBSERVATION: THE ADVANTAGE OF "REAL TIME" DATA ACQUISITION

Albert R. Hollenbeck and Ronald G. Slaby
University of Washington

Abstract

Two observers who were using an electronic digital data acquisition system were spot checked for reliability at random times over a four month period. Between- and within-observer reliability was assessed for frequency, duration, and duration-per-event measures of four infant behaviors. The results confirmed the problem of observer drift--the fluctuations of scores across sessions--for the frequency and duration-per-event measures. In contrast, the "real time" duration scores were stable across sessions, indicating the robustness of this measure of behavior.

TM004 264

MEASURES OF RELIABILITY IN BEHAVIORAL OBSERVATION: THE ADVANTAGE OF "REAL TIME" DATA ACQUISITION[1]

Albert M. Hollenbeck[2] and Ronald G. Slaby[3]
University of Washington

A number of recent studies have been addressed to the problem of assessing reliability in observational research (e.g., Reid, 1970; Johnson and Bolstad, 1973; Whalen, 1974). This interest in the more subtle aspects of the observational process reflects a general increase in the use of direct observational measures in psychological research. Reliability in observational research generally means that two or more observers independently record the same naturally occurring behavioral events in a similar way. It has been generally assumed that if two observers achieve a high level of inter-observer reliability, then they are measuring the same aspects of behavior across sessions as well. Yet, Reid (1970) has demonstrated that this may not be the case. His data indicate that whereas reliability between observers can remain at a constant level, reliability across sessions for a given observer tends to decrease with time. This problem of observer drift, i.e., the fluctuation of observations across sessions, has important implications both for the interpretations of data already collected and for the collection of data in future research.

A second problem in establishing reliability is that of selecting an appropriate statistical index. The typical measure of reliability has been some summary statistic such as a percentage agreement score or a correlation coefficient. It has been generally assumed that these traditional statistical measures used to compute reliability are valid. However, percentage agreement and correlation measures of reliability have recently come under justifiable criticism (e.g., Hartmann, 1974) for over-estimating

reliability as well as for being insensitive to the detection of chance agreements.

In order to overcome these difficulties in establishing reliability, systematic attempts have been made to evaluate different components of the observational process (e.g., Heath, 1973; Tuplin and Reid, 1973; Kazdin, Brown, and LaBlanc, 1973). These evaluations have uncovered several questions underlying the basic assumptions of the behavioral observation methods. For example, in the majority of observational studies, the most common measure used has been the frequency of occurrence of some behavioral event. In fact, most studies have been time-sampled in such a way that actual frequencies are not scored. Rather, a modified-frequency score (i.e., a score based on the number of arbitrarily defined time intervals in which an event has occurred) is used to mark simple occurrence or non-occurrence of a behavioral event. The very nature of modified-frequency measurement is suspect, since actual frequencies and durations are confounded.

Recent advances in technology have provided electronic systems which allow the unconfounded recording of actual frequency and duration scores. MEDICAS and the Behavioral Observation Scoring System (BOSS) are two such systems (Herbert, Stephenson, and Rappenthal, 1973). These advances, which allow the experimenter to measure exact frequencies and durations separately, raise several interesting questions. How does the reliability of real frequency and real duration measures compare with that of the modified-frequency measures typically used in previous research? Are the unconfounded frequency and duration measures subject to fluctuations in observations across sessions, as reported by Reid and

observe the modified-frequency account. The[illegible] purpose of this study was to evaluate the reliability of observations gathered on both frequency and duration measures. This study was designed to answer questions with to the reliability of these measures.

Method

Subjects

Two female undergraduates at the University of Washington served as observers in this study. Both observers at both volunteered for received academic credit for applied field work as a psychology.

Apparatus

A videotape machine was used to record the responses of a storewide-old infant to a stimulus presentation designed to elicit reactions from the infant. This videotaped sequence was presented to the observers for purposes of assessing observer reliabilities.

The behavior code was a modified version of the one previously used by Hollenbeck (1972). Five mutually exclusive and exhaustive behavioral categories--Vocalization, Head Movement, & Gaze Movement, Body Movement, and No Behavior--were hierarchically arranged (I did arrange as a priority) terms. Specifically, Vocalizations made by the infant took overlap preference over Head Movements when both behaviors occurred simultaneously. In the same way, Head Movements were scored over & Gaze Movements, Gaze Movements over Body Movements; and any movements at & resolutions took overlap preference over the No Behavior category.

The behavioral Observation Recording & System (BORS) was used to record the coded data. BORS is an electronic multifield data acquisition system

[illegible]

must be considered. Based on the findings of the present study, duration measures appear to be less suceptible to observer drift. In addition, assessment of reliability through analyses of variance and regression analyses should be further explored, considering the potential advantages in precision and sensitivity.

References

Cohen, Jacob. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70, 426-443.

Hartmann, Donald P. Assessing the quality of observational data. Paper presented at the Meeting of the Western Psychological Association, San Francisco, April, 1974.

Hawn, Joyce, Brown, George and LeBlanc, Judith M. A comparison of three observation procedures: consecutive intervals on-the-spot; consecutive intervals from video tape; 10-sec-on, 10-sec-off from video tape. Paper presented at the Meeting of the American Psychological Association, Montreal, August, 1973.

Hollenbeck, A. R. Imitation as a skill in infancy. Unpublished manuscript, University of Washington, Seattle, 1972.

Johnson, S. M., and Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Champaign, Ill.: Research Press, 1973.

Mash, Eric J., and McElwee, John D. Situational effects of observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. Child Development, in press.

Reid, John B. Reliability assessment of observation data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Sackett, G. P., Stephenson, E., and Ruppenthal, G. C. Digital data acquisition systems for observing behavior in laboratory and field settings. Behavioral Research Methodology and Instrumentation, 1973, 5, 344-348.

Taplin, P. S., and Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. Child Development, 1973, 44, 547-554.

Whelen, B. Reliability of Human Observers. Unpublished doctoral dissertation, University of Utah, Salt Lake City, Utah, June, 1974.

Footnotes

1. An earlier version of this paper was presented at the 82nd Annual Meeting of the American Psychological Association, Division 25, August 30-September 3, 1974, New Orleans, Louisiana. This research was supported, in part, by the Child Development and Mental Retardation Center of the University of Washington.

2. Requests for reprints should be sent to Albert E. Hollenbeck, Department of Psychology, NI-25, University of Washington, Seattle, Washington 98195.

3. The authors wish to express appreciation to Gene Sackett for his guidance throughout all phases of this research, and to Beverly Davis and Lynn Davis for their work as observers in this study.

Table 1

Analysis of Variance for Frequency

Source	df	MS	F
Total	13	1420.51	
Residual (error)	37	5.04	
Observer (O)	1	7.22	ns
Vocalizations (V)	1	8217.42	1530.48***
Head Movement (H)	1	28.03	5.50**
Arm Movements (A)	1	1520.87	301.65**
Body Movement (B)	1	6993.80	1387.66***
Trends			
Linear Trend (T)	1	153.76	30.51**
Quartic Trend	1	28.28	5.61**
Cubic Trend[a]			
Quadratic Trend[a]			
O X T	1	12.96	ns
O X V	1	36.93	7.36***
O X H	1	20.83	4.13*
O X A	1	6.67	ns
O X B	1	20.80	7.97*

[a]The quadratic and cubic trends were eliminated from the analysis by t he computer progres due to the small SS attributable to these factors.

*p < .05

**p < .025

***p < .001

Table 2

Analysis of Variance for
Durations-per-Event

Source	df	MS	F
Total	34	421.41	
Residual (error)	25	5.56	
Observer (O)	1	121.68	21.88***
Vocalizations (V)	1	1375.28	248.34***
Head Movement (H)	1	961.33	173.26***
Arm Movement (A)	1	5218.37	940.83***
Body Movement (B)	1	2690.98	485.67***
Trends			
Linear Trend (T)	1	154.33	27.77***
Quadratic Trend	1	30.19	5.43*
Cubic Trend	1	1.88	ns
Quartic Trend	1	14.05	ns
O X T	1	13.69	ns
O X V	1	30.42	5.47*
O X H	1	.03	ns
O X A	1	1.47	ns
O X B	1	33.80	6.10**

*p < .05

**p < .025

***p < .001

Tabela 5.7

Analysis of Variance Table

Source	df	SS	F
Total	59	[illegible]	
Residual (error)	50	[illegible]	
Block	1	[illegible]	NS
Fert	1	[illegible]	[illegible]
Fert	1	[illegible]	[illegible]
Fert	1	[illegible]	[illegible]
Fert	1	[illegible]	[illegible]
Trees			
[illegible]	1	[illegible]	NS
[illegible]	1	[illegible]	NS
[illegible]	1	[illegible]	NS
[illegible]	1	[illegible]	NS
[illegible]	1	[illegible]	NS
[illegible]	1	[illegible]	[illegible]
[illegible]	1	[illegible]	[illegible]
[illegible]	1	[illegible]	[illegible]
[illegible]	1	[illegible]	[illegible]

$^{a}_{g}$ = .05

$^{**}g$ = .01

$^{***}g$ = .001

Figure caption

Fig. 1. Nonparametric differences of posterior means (bias) in the four strata subjected (box) the grand mean for Poisson and binomial errors.

Percentage Difference of Session Means from the Grand Mean